

Práctica: Análisis de datos para la promoción de la salud

Presentación

Prevención de la sarcopenia basada en datos

La sarcopenia es la pérdida de masa muscular debido al proceso de envejecimiento. Es, por lo tanto, un proceso natural ya que a partir de los 40 años empezamos a perder masa muscular a un ritmo del 3-8% por década. Sin embargo, la pérdida de masa muscular se acelera a partir de los 60 años, dando lugar en última instancia a tal disminución de la fuerza física que dificulta las actividades cotidianas como levantar peso, subir escaleras o incluso caminar. La sarcopenia lleva irremediablemente a una vida aún más sedentaria, lo cual complica el cuadro clínico del paciente por la aparición de problemas típicos de la discapacidad y el sedentarismo, tales como la diabetes. Además, la disminución de la calidad de vida asociada a la sarcopenia y la dependencia de otras personas hace que esta enfermedad constituya un problema de salud pública de primer orden. Se estima que en poblaciones no asiáticas la incidencia de la sarcopenia es del ~20% en la población adulta (Shafiee *et al*, 2017).

Causas, síntomas y tratamiento

Las causas de la sarcopenia tienen que ver con la disminución natural y progresiva de nuestra capacidad de producir proteínas y hormonas, tales como la testosterona y el insulín-like growth factor 1 (IGF-1).

Los síntomas más frecuentes de la sarcopenia son:

- Pérdida de masa muscular
- Debilidad física
- Pérdida de la capacidad de resistencia
- Pérdida de equilibrio y dificultad para subir escaleras

Estos síntomas, más que síntomas, son una confirmación del diagnóstico de la enfermedad. Se conocen algunos factores de riesgo, tales como un estilo de vida sedentario y una mala nutrición, pero de nuevo son demasiado generales como para poder hacer predicciones fiables.

En la actualidad no existe ningún tratamiento médico contra la sarcopenia, aunque se están investigando terapias hormonales (testosterona y somatotropina) para ayudar a aumentar la masa muscular. La única solución conocida contra la sarcopenia es la detección temprana junto con la estimulación física y el seguimiento de una dieta adecuada.

Por lo tanto, la investigación de diversas combinaciones más amplias de síntomas sería muy importante para intentar predecir la sarcopenia lo antes posible y así poner a los pacientes de riesgo en programas que retrasen la aparición de la enfermedad. Además, cuando la sarcopenia se encuentra asociada a otras enfermedades típicas del envejecimiento sus efectos pueden ser aún más pronunciados. Por ejemplo, cuando se juntan la sarcopenia y la osteoporosis, el riesgo de caídas y fracturas es mucho más alto. Por lo tanto, la detección temprana de la sarcopenia y su tratamiento ayudará a su vez a disminuir el efecto de otras enfermedades asociadas a la vejez.

Objetivos

El objetivo de esta actividad es asimilar y poner en práctica los conocimientos adquiridos desde principios de semestre, y en este módulo en particular, en cuanto a:

- Programación en R
- La obtención de valor a partir del análisis de los datos
- La evaluación de un modelo de análisis de datos
- Visualización de datos
- Hacer recomendaciones prácticas en el contexto particular de la salud pública

Enunciado

El objetivo de esta práctica es analizar una serie de datos históricos de pacientes de sarcopenia para determinar si existen una serie de síntomas más allá de los ya conocidos y que puedan ayudar a predecir la enfermedad antes de que se diagnostique. No todas las personas desarrollan un grado de sarcopenia que les inhabilite, y de las que la desarrollan, no todas lo hacen a la vez.

Se ha hecho un seguimiento detallado de los síntomas que presentaban miles de pacientes de sarcopenia. La población en cuestión es el país de Atlantis, y los datos han sido recopilados durante un espacio de 25 años. El contexto se presenta a continuación, así como las preguntas específicas que deben responderse mediante el análisis de los datos.

Atlantis

Atlantis es un país insular situado en el océano Atlántico, a unas 1.500 millas náuticas al oeste de las Islas Canarias. Atlantis tiene una población de 1,2 millones que se reparte geográficamente en un archipiélago de 11 islas que rodean a la isla principal (Ho'okipa). El principal problema médico de Atlantis es el envejecimiento de su población: más de la mitad de los atlantianos tiene más de 70 años, y de éstos un 35% presenta problemas derivados de la sarcopenia. Ésto hace que la atención geriátrica de una parte importante de la población

suponga un gasto inabordable, además de no haber suficientes hogares para mayores ni personal especializado.

Por lo tanto, un objetivo principal es implantar una estrategia de medicina preventiva que ayude a anticipar los casos de sarcopenia para así ayudar a retrasar la degeneración muscular de los pacientes. De esta forma la población de mayores será más activa y su gestión hospitalaria será más factible.

Los síntomas conocidos de la sarcopenia son de poco valor predictivo porque no sólo son bastante generales sino que más que síntomas, son una confirmación de la enfermedad. En el mundo tampoco hay demasiados estudios exhaustivos de síntomas asociados a la sarcopenia. En Atlantis, sin embargo, la doctora Ellen Ripley lleva 25 años recopilando esta información en anticipación del envejecimiento de la población y por el hecho de que los atlantianos lleven 450 años aislados genéticamente de los primeros colonos europeos. Por lo tanto, es posible que presenten una serie de síntomas particulares producto de su genética y del medio en el que viven.

Objetivo de la práctica

Eres un *data scientist* a quien han encargado la importante tarea de determinar los principales síntomas de la sarcopenia en la población de Atlantis. Contáis con los registros de la doctora Ellen Ripley, que consisten en la anotación de los síntomas de la sarcopenia en 9430 pacientes diferentes.

La estrategia sugerida es la de realizar un análisis de reglas de asociación para determinar los síntomas más característicos de la sarcopenia en esta población. Las preguntas a responder son las siguientes:

1. Explica qué son las reglas de asociación y proporciona algunos ejemplos de aplicación concretos, tanto en el contexto médico como otro más general. Los ejemplos deben estar apropiadamente referenciados **(1 punto; máx. 1 pág.)**
2. Caracterización del dataset: ¿Explica cómo has caracterizado los datos erróneos que deben limpiarse antes de seguir con el análisis y por qué crees que son erróneos? ¿Cuántos casos médicos y cuántos síntomas diferentes hay en el dataset? ¿Cuántos pacientes presentan uno o más síntomas? ¿Cuáles son los 5 síntomas más comunes? Detalla todas las caracterizaciones descriptivas de este dataset que te parezcan relevantes.

Finalmente, es preciso explicar qué conclusiones preliminares se pueden obtener de este análisis descriptivo y explicar cómo se podría intentar validar estas conclusiones. Es importante también utilizar técnicas de visualización de datos adecuadas (todas las que se quieran). **(1 punto; máx. 3 pág.)**

3. Diseña una estrategia optimizada para obtener reglas de asociación a partir de este dataset. Es preciso describir el razonamiento detalladamente. **(2 puntos; sin máximo de páginas, pero es necesario ser escueto).**
4. Conjunto final de reglas: ¿Cuántas reglas obtuviste, qué te indican, y cómo las representarías visualmente? **(1 punto; máx. 1 pág.)**
5. Diseña una estrategia basada en el dato para la gestión de la sarcopenia en Atlantis. Debes explicar asimismo cómo se integraría esta estrategia en un sistema sanitario de información **(3 puntos; máx. 2 pág.)**
6. Escribe un pequeño ensayo sobre las limitaciones que tienen las reglas de asociación como herramienta de *data mining*. ¿Las reglas de asociación tienen carácter predictivo (razona la respuesta)? **(2 puntos; máx. 1 pág.)**

Materiales

Referencias generales por orden de importancia

- Paquete de **arules** y ejemplos trabajados: **Introduction to arules – A computational environment for mining association rules and frequent item sets.**

URL: <https://cran.r-project.org/web/packages/arules/vignettes/arules.pdf>

- Agrawal, R., Imieliński, T. and A. Swami (1993) **Mining association rules between sets of items in large databases.** *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93.* p. 207. ISBN 0897915925. doi:10.1145/170035.170072.

Ésta es la publicación original del algoritmo de Association Rule Mining. Está disponible a través de la biblioteca de la UOC.

- Shafiee, G., Keshtkar, A., Soltani, A., Ahadi, Z., Larijani, B., Heshma, R. (2017) **Prevalence of sarcopenia in the world: a systematic review and meta- analysis of general population studies.** *Journal of Diabetes & Metabolic Disorders* vol. 16, p. 21.

URL: <https://jdmdonline.biomedcentral.com/articles/10.1186/s40200-017-0302-x>

Éste es una revisión general sobre la sarcopenia.

- Fournier-Viger P., Lin J.C-W., Vo B., Chi T.T., Zhang J. and H.B. Le (2017) **A survey of itemset mining**. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery vol. 7(4): e1207.

Ésta es una revisión crítica sobre las diversas técnicas de *association rule mining* y sus aplicaciones.

Ejemplos prácticos

Algunos ejemplos que os pueden ayudar a solucionar la prueba de evaluación continua planteada son:

- A Gentle Introduction on Market Basket Analysis—Association Rules

URL:

<https://www.r-bloggers.com/a-gentle-introduction-on-market-basket-analysis%E2%80%828A-%E2%80%828Aassociation-rules/>

- Market Basket Analysis with R

URL: <http://www.salemmarafi.com/code/market-basket-analysis-with-r/>

Criterios de evaluación

Para evaluar esta actividad se tendrá en cuenta:

- La destreza técnica y corrección del análisis de datos.
- La argumentación y el razonamiento en las respuestas.
- El soporte de bibliografía externa para justificar los argumentos presentados.
- La presentación del formato del contenido y responder dentro de los límites establecidos.

Formato de entrega

Se deberá entregar un archivo en formato PDF con las respuestas a las preguntas de esta práctica, mezclando código de R y resultados, así como cualquier otro texto.

La forma más **rápida, fácil, elegante, profesional y reproducible** es trabajar con RStudio donde tienes inmediatamente una interfaz donde introducir el código R y analizar los resultados. Luego para producir un documento final que integre el código, los resultados y cualquier otro texto que queráis añadir (por ejemplo para las respuestas que son solo de texto), entonces dentro de RStudio puedes abrir un proyecto (**File -> New File -> R Markdown**), y allí veréis una interfaz donde te permite separar código R de texto normal. Por ejemplo, el primer bloque de código R que tengo para la respuesta modelo es el siguiente:

```
```{r bloque 1}

library(arules)

trans <- read.transactions("data.txt",format="single",
cols=c(1,2),sep=" ",header=TRUE,rm.duplicates=TRUE)

trans
```
```

Luego, cualquier texto que vaya fuera de ````{r bloque 1}` y de las ````` que lo cierran es considerado texto normal. Tened en cuenta que el segundo bloque de código R hay que llamarlo de otra forma porque si no RStudio se confunde (podéis llamarlo “bloque 2” por ejemplo).

Finalmente, y esto es lo más importante del fichero tipo R Markdown, es que cuando queráis producir el documento final podéis hacer en la misma interfaz **Knit -> Knit to PDF** y con esa única instrucción se generará un PDF donde combinará el código, los resultados integrados después del código, y también cualquier otro texto que hayáis incluido. Esto sólo funcionará, sin embargo, si el RStudio puede correr todo el código incluido dentro del documento R Markdown, de principio a fin.

Puede pasar que, por lo que sea, no exista una opción directa a PDF (depende de la versión de RStudio y del ordenador). También podéis enviarlo en **HTML (knit to HTML)**. Funciona igual.

También, hay otra manera. Una forma más larga y pesada de hacerlo es ejecutar el código de R, y luego pegarlo junto con los resultados en un documento **Word**, el cual al final acabará convertido en PDF.

En la wiki de R que hay colgada en este aula tenemos un video sobre la instalación de RStudio y los primeros pasos.